Skin Lesion Classification With Deep CNN Ensembles

Sara Atito Ali Ahmed and Berrin Yanıkoğlu
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey
saraatito,berrin@sabanciuniv.edu

Özgü Göksu and Erchan Aptoula
Institute of Information Technologies
Gebze Technical University
Kocaeli, Turkey
ogoksu,eaptoula@gtu.edu.tr

***Abstract*—Early detection of skin cancer is vital when treatment is most likely to be successful. However, diagnosis of skin lesions is a very challenging task due to the similarities between lesions in terms of appearance, location, color, and size. We present a deep learning method for skin lesion classification by fusing and fine-tuning three pre-trained deep learning architectures (Xception, Inception-ResNet-V2, and NasNetLarge) using training images provided by ISIC2019 organizers. Additionally, the outliers and the heavy class imbalance are addressed to further enhance the classification of the lesion. The experimental results show that the proposed framework obtained promising results that are comparable with the ISIC2019 challenge leader board.**

***Keywords*—Skin Lesion Classification, ISIC, Deep Learning, Convolutional Neural Networks, Anomaly Detection, Ensemble**

I. Introduction

Computer-aided diagnostic tools have long empowered pathologists against a wide spectrum of diseases, since the first development of expert systems in the 1970s. Their performance levels have nowadays reached unprecedented levels, mostly thanks to the paradigm shifting advances in the field of machine learning, skin cancer in particular, as the most common form of this often fatal disease, has received particular attention in this regard and deep learning methods have reached a level of precision that is comparable to qualified dermatologists.

As with most diseases, early diagnosis of a particular strain of cancer is of crucial significance for the patient's successful treatment. Even though a human expert can be trained to achieve a diagnostic accuracy of skin cancer types up to approximately 80% [1], the number of dermatologists is unfortunately insufficient when compared against the disease occurrence frequency [2].

In an effort to rectify this imbalance, the International Skin Imaging Collaboration (ISIC) has developed the ISIC Archive, an international repository of validated dermoscopic images around which the ISIC challenge has been organized annually, in order to boost the development and effectiveness of appropriate computer-aided diagnostic tools.

As expected, the ISIC challenge is becoming progressively harder and more akin to real-world scenarios. This year, instead of segmentation and attribute detection tasks, the entire challenge focuses on lesion diagnosis. The dataset contains 8 strains of skin cancer (one more than 2018). Besides, the diagnostic objective has been upgraded to include a "None of the others" class as well, rendering it as an open set recognition problem. Random samples from the ISIC2019 dataset are shown in Figure 1.

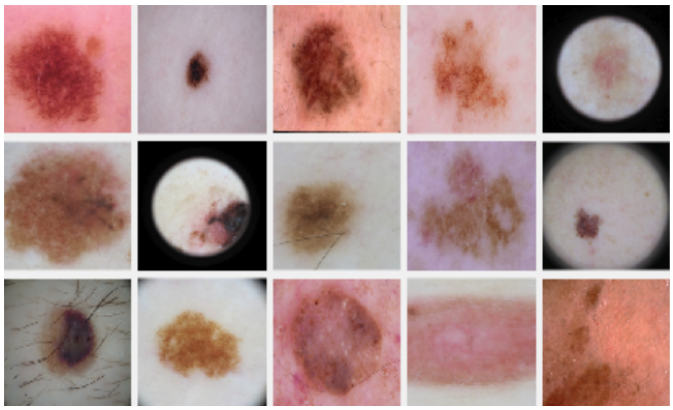

Figure 1: Random samples of skin lesions from ISIC2019 Training set.

This paper presents the developed system for the ISIC2019 challenge, and details our findings. Our system relies on an ensemble of various modern convolutional neural networks varying from each other in terms of architecture, preprocessing and data augmentation techniques. Furthermore, a comprehensive study of fusion strategies has been conducted, further supported by state of the art gradient boosting methods. Finally, special precautions have been taken for anomaly detection, so as to handle the case of samples stemming from unknown classes. Our proposed method obtained promising results that are comparable with the ISIC2019 challenge leader board [1].

The rest of this paper is organized as follows: Section II describes the developed system based on the fine-tuning of Xception, Inception-ResNet-V2, and NasNetLarge models for skin lesion classification. Next, Section III is dedicated to the description of the utilized dataset, data augmentation, and classifiers' fusion and presentation of designed experiments and their results. The paper concludes in Section IV with a summary and discussion of the utilized methods and obtained results.

[1] https://challenge2019.isic-archive.com/leaderboard.html

II. Skin Lesion Classification

In recent years, there have been many breakthroughs in the development of deep learning using Convolutional Neural Networks (CNN). In this work, we tackled the skin lesion classification problem using three of the latest and most accurate models, namely Xception [3], Inception-ResNet-V2 [4], and NasNetLarge [5].

Xception [3] is an extreme version and an extension of the Inception [6] architecture, which replaces the standard Inception modules with depth-wise separable convolutions. The network is 71 layers deep with only 22.9 million parameters and an image input size of 299-by-299. Inception-ResNet-V2 [4] is an advanced convolutional neural network that combines the inception module with ResNet [7] to increase the efficiency of the network. As for NasNetLarge [5], authors propose to search for an architectural building block on a small dataset and then transfer the block to a larger dataset. Initially, they search for the best convolutional layer on CIFAR-10, then apply this layer to ImageNet by stacking together more copies of this layer. They also proposed a new regularization technique called ScheduledDropPath that significantly improves the generalization of their proposed network.

Our approach is based on fine-tuning and fusing of the aforementioned three successful deep learning models. These three models are currently the top-ranked architectures of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014. The models are pre-trained on the ILSVRC 2012 dataset with 1.2 million labeled images of 1,000 object classes. Different network configurations are used to further handle the class imbalance. The distribution of the eight given categories of the ISIC dataset is shown in Figure 2.

Ensemble learning techniques have seen a huge jump in popularity in the last years. Ensemble can help in building a much robust model from a few weak models, which eliminates a lot of the model tuning that would otherwise be needed to achieve good results. In this work, we used LightGBM [8], one of the most famous ensemble techniques nowadays.

LightGBM is an open-source framework which trains a Gradient Boosted Decision Tree (GBDT). In GBDT, successive models are found by applying gradient descent in the direction of the average gradient, calculated with respect to the error residuals of the loss function of the leaf nodes of previous models. In this work, we trained LightGBM using the extracted features from the last pooling layer of our trained models.

All training and testing were conducted on a Linux system with a Titan X Pascal GPU and 12GB of video memory.

A. Anomaly Detection

The goal of the ISIC2019 competition is to classify dermoscopic images among nine different diagnostic categories while only eight classes are given for training. One way of dealing with the unknown class is to consider all of the instances coming from this class as outliers and target them using one-class learning approaches. One-class learning is a challenging task especially when dealing with high dimensional data points. In this paper, we applied one-class learning using deep neural network features and compared classifier performance based on the approaches of OC-SVM [9], Isolation Forest [10], and Gaussian Mixtures [11] as shown in Section III. We found that the best approach for this dataset is Isolation Forest [10]. Isolation Forest is based on the fact that the features of anomalies are very different from the normal samples. The idea is to build an ensemble of isolation trees where anomalies have short average path lengths on the those trees.

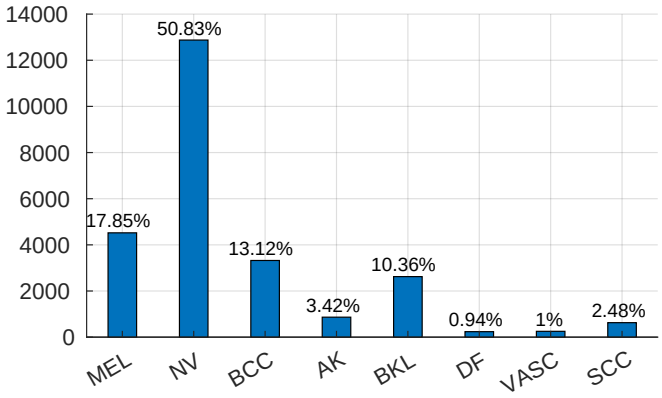

Figure 2: Distribution of the available ISIC2019 training images across the eight given skin lesion categories.

III. Experiments and Results

The training data of ISIC2019 includes skin lesion images from several datasets, such as: HAM10000 [12], BCN20000 [13], and MSK [14] datasets. The goal of ISIC2019 is to classify dermoscopic images among nine different diagnostic categories: 1. Melanoma (MEL); 2. Melanocytic nevus (NV); 3. Basal cell carcinoma (BCC); 4. Actinic keratosis (AK); 5. Benign keratosis (BKL); 6. Dermatofibroma (DF); 7. Vascular lesion (VASC); 8. Squamous cell carcinoma (SCC); and 9. None of the others (UNK). The dataset consists of $25,331$ images for training across 8 different categories. Furthermore, the test dataset contains an additional outlier class that is not represented in the training data.

Two tasks are available for this competition: 1) classify dermoscopic images without meta-data, and 2) classify images with additional available meta-data. In this paper, we target the first task where only the provided images are used without any usage of meta-data or external dataset.

A. Ensemble of Deep Neural Networks

Generally, neural networks have high variance due to the stochastic training approach that make them sensitive to the nature of the training data. The models may find a different set of weights each time they are trained, which in turn may produce different predictions.

A successful approach to reduce the variance of neural network models is ensemble learning, where multiple models are trained instead of a single model and then combining the predictions from these models. Not only this approach reduces the variance of the predictions but also can result in predictions that are better than any single model.

Therefore, we trained several convolutional neural network models to tackle this problem. At first, we split the training

Architecture	Specifications		
	Batch Size	**# of Epochs**	**Loss Function**
Xception	32	200	Cross Entropy
	32	30	Focal Loss with $\gamma = 1$
	32	30	Focal Loss with $\gamma = 2$
	32	80	Focal Loss with $\gamma = 3$
	32	30	Focal Loss with $\gamma = 4$
Inception-ResNet-V2	20	50	Cross Entropy
	32	70	Cross Entropy
	64	90	Cross Entropy
NasNetLarge	20	25	Cross Entropy

Table II: Specifications of the trained CNN models

set into 80-20% ratio to create the validation set to fine-tune the learning rate. We found that the best learning rate for the three used models, Xception, Inception-ResNet-V2, and NasNetLarge is 0.01 with validation accuracy around 90%.

We implemented Xception, Inception-ResNet-V2, and NasNetLarge models using Matlab's Deep Learning Toolbox. All the weights were fine-tuned from the pre-trained weights on the ImageNet dataset, while the last layer was learned from scratch. We used the same learning rate (0.01) for all of the systems.

During training, several data augmentation techniques were applied, such as heavy rotation $[-90$ to $90]$, x and y translation $[-10$ to $10]$, vertical and horizontal flipping. All data augmentation were applied on the fly, which means, at every iteration, different setting of augmentations are applied on top of the original batch of images.

The specifications of the trained models are shown in Table II where our systems are trained with different batch sizes and different number of epochs. Also, we employed different loss functions, namely, cross entropy and focal loss. Focal loss function (Equation 1) is used to address the imbalance between classes. Various Xception networks were trained with α_i set to the inverse class frequency and several values of γ_i as shown in Table II.

$$FL(p, y) = -\sum_i \alpha_i y_i (1 - p_i)^\gamma \log(p_i) \qquad (1)$$

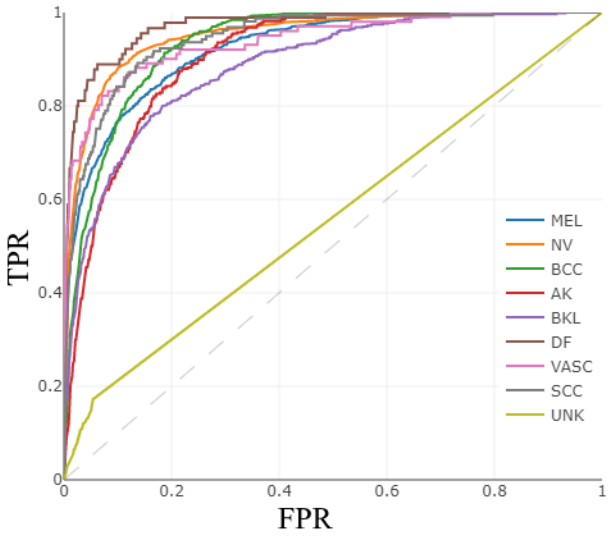

Figure 3: ROC curve of the nine skin lesion categories using deep CNNs ensembles.

where p_i and y_i are the prediction and the ground-truth of a given sample, respectively.

Finally, in testing time, we applied test time augmentation (TTA). Specifically, we applied rotation with 90, 180, 270 degrees with and without horizontal flipping to have 6 augmented images. In addition, we applied 30 random augmentations similar to the techniques applied during training but with a smaller rotation range: [-15, 15]. To further boost the efficiency and reduce the variance, we trained a LightGBM module using the extracted features of the last pooling layer of each trained model. Score-level averaging is applied to combine the prediction scores assigned to each class for all the augmented patches; locally, within a single network and globally, among different models. The probability of the UNK class is set to $1 - max$ of the probabilities of the other eight classes for a given sample.

Figure 3 shows the ROC curve that is plotted with true positive rate against the false positive rate of each lesion category, individually. Table I shows the performance comparison of our model to the top two ranking results in the ISIC2019 challenge leader board of the eight given classes of skin lesion classification task. It shows that our approach surpassed top-2 rank with a high margin equals to 0.1246 and achieved comparable results with the top-1 rank method, despite usage of external data.

As for the unknown class in the ISIC dataset, we addressed it by the notion of anomaly detection. We applied one class

Method	External Data?	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	**Mean**
		AUC (Area Under the Curve)								
Top-1 Rank	Yes	0.928	0.960	0.949	0.914	0.904	0.979	0.956	0.938	0.9410
Top-2 Rank	No	0.808	0.878	0.868	0.765	0.762	0.832	0.797	0.744	0.8067
Ours	No	0.925	0.951	0.934	0.902	0.885	0.968	0.941	0.944	0.9313
		Accuracy								
Top-1 Rank	Yes	0.900	0.889	0.912	0.940	0.934	0.987	0.986	0.975	0.9404
Top-2 Rank	No	0.896	0.902	0.888	0.916	0.927	0.982	0.984	0.962	0.9321
Ours	No	0.903	0.894	0.873	0.945	0.923	0.989	0.989	0.980	0.9370

Table I: Performance of our model to the top two ranking results on ISIC2019 leader board.

Anomaly Class	Validation Accuracy	Precision	Recall
Class 1	92.76%	94.72%	89.88%
Class 2	97.16%	88.13%	90.07%
Class 3	94.08%	98.14%	89.81%
Class 4	90.01%	97.38%	90.08%
Class 5	91.26%	94.25%	89.97%
Class 6	90.64%	99.77%	90.26%
Class 7	90.78%	100%	90.16%
Class 8	90.05%	98.67%	89.58%

Table III: Performance of Isolation Forest using deep learning features extracted from the last pooling layer of one of the trained Xception network.

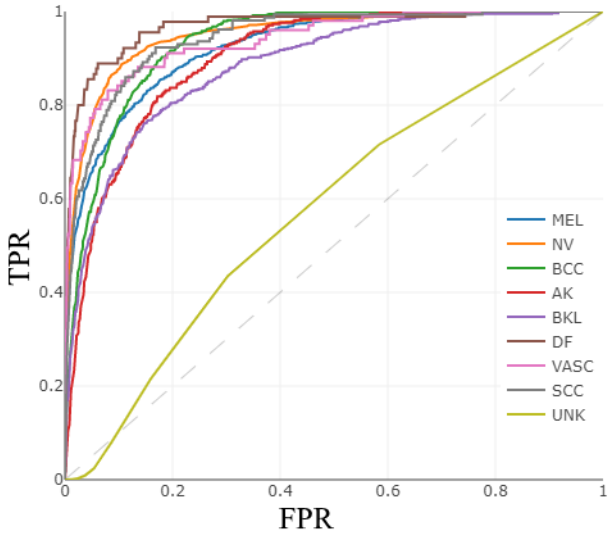

Figure 4: ROC curve of the nine skin lesion categories after incorporating anomaly detection.

learning using the features of the last pooling layer from the trained networks as they are considered to be more powerful representations of the images than handcrafted features.

We tried several one-class learning approaches like one-class support vector machines (OC-SVM), Isolation Forest and Gaussian Mixtures. We found that Isolation Forest is the best approach to be used in this task.

To show the empirical effectiveness of this step, iteratively, one class from the eight given classes is chosen to be an outlier and removed from the training procedure. In other words, for each experiment, we set the validation set to be all of the samples belonging to the anomaly class in addition to 20% from the other classes and the rest are left for training. The performance of Isolation Forest is shown in Table III. To incorporate isolation forest into our approach, we used the features of the whole training set coming from the last pooling layer of our trained models as an input to isolation forest. In testing time, we assigned the probability of the UNK class to the probability coming from isolation forest. As shown in Figure 4, The ROC curve of the UNK class indicates the effectiveness of adding anomaly detection to our approach compared to the ROC curve of UNK class in Figure 3 to improve the prediction of the UNK class.

IV. Conclusions

The core of our approach is based on an ensemble and fusing of three pre-trained deep learning architectures (Xception, Inception-ResNet-V2, and NasNetLarge) using training images provided by the ISIC2019 organizers. LightGBM and one class classification models are used to further boost our predictions. In future work, we would like to investigate deep learning approaches for anomaly detection.

Kaynaklar

[1] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The lancet oncology*, vol. 3, pp. 159–165, 2002.

[2] A. Kimball and J. R. Jr, "The US dermatology workforce: a specialty remains in shortage," *Journal of the American Academy of Dermatology*, vol. 59, pp. 741–745, 2008.

[3] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[5] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[9] B. Scholköpf and A. Smola, "Support vector machines, regularization, optimization, and beyond," *Learning with Kernels*, 2002.

[10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[11] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 155–176, 1996.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[13] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A. C. Halpern, S. Puig, and J. Malvehy, "Bcn20000: Dermoscopic lesions in the wild," 2019.

[14] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kallo, K. Liopyris, N. M. H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 168–172.